跨语言情境下基于对抗的实体关系抽取模型研究*

余传明[1] 王曼怡[2] 安璐[3]
[1] 中南财经政法大学信息与安全工程学院 武汉 430073 [2] 中南财经政法大学统计与数学学院 武汉 430073
[3] 武汉大学信息管理学院 武汉 430072

摘 要：[目的/意义]从实体关系抽取视角出发，将单一语言情境下的知识获取任务扩展到跨语言情境，提升低资源语言的关系抽取效果。[方法/过程]提出一种跨语言对抗关系抽取（Cross-Lingual Adversarial Relation Extraction，CLARE）框架，将跨语言关系抽取分解为平行语料获取和对抗适应关系抽取两个子模块。通过词典扩展或自学习方法将源语言关系抽取数据集转换为目标语言数据集，在此基础上利用对抗特征适应将源语言的特征表示迁移给目标语言，再利用训练得到的目标语言关系抽取网络对目标语言进行关系分类。[结果/结论]将本文方法应用到以 ACE2005 多语言数据集为基础的英语－中文、中文－英文两种跨语言关系抽取任务上，最优模型的 Macro-F1 值分别为 0.880 1 和 0.842 2。实验结果表明本文提出的跨语言对抗关系抽取 CLARE 框架能显著提升低资源语言实体关系抽取的效果。研究结果对于改进跨语言情境下的关系抽取模型以及促进实体关系抽取研究在情报学领域的应用具有重要意义。
关键词：跨语言信息抽取 实体关系抽取 深度学习 生成对抗网络
分类号：TP391
DOI：10.13266/j.issn.0252－3116.2020.17.014

1 引言

互联网技术发展日新月异，人们需要处理的数据量激增，领域交叉现象越来越突出，如何快速高效地从这些开放领域的文本中抽取出有效信息，成为摆在人们面前的重要问题。实体关系抽取[1]（Entity Relation Extraction，ERE），又称为关系抽取（Relation Extraction，RE），是指通过对文本信息建模，自动抽取出句子中实体对之间的语义关系，提取有效的语义知识。例如，句子“Bill_Gates is the founder of Microsoft.”中包含一个实体对（Bill_Gates，Microsoft.），这两个实体对之间的关系为 Founder。实体关系抽取被广泛应用于文本摘要[2]、自动问答[3]、机器翻译、语义标注以及知识图谱构建[4]等任务中。

过去大多数关系抽取模型关注单语言数据（以标注资源丰富的英文文本为主），而对于标注语料相对稀缺的语言（如日语、法语等），由于手工标注数据集获取昂贵且费时、远程监督数据集标注噪音难以排除，难以建立有效的关系抽取模型。研究构建跨语言关系抽取的意义在于：①由于各语种知识分布不均匀，通过数据集扩展可以有效地弥补目标语言数据集的不足，实现低资源语言的实体关系抽取；②可以充分利用多语种在知识表达方式上的互补性，增加知识的覆盖率和共享度。跨语言关系抽取可以应用于跨语言的信息检索、机器翻译[5]、知识问答以及跨语言知识图谱的构建等任务中。由于其广泛的应用前景，跨语言实体关系抽取正得到学术界及工业界的广泛重视。

在上述背景下，本文将跨语言平行语料获取与跨语言关系抽取任务相结合，提出一种跨语言情境下基于生成对抗的实体关系抽取框架，并将其应用于“源语言英语－目标语言中文”和“源语言中文－目标语言英语”两种不同的跨语言任务，以期检验模型对于提升目标语言实体关系抽取的效果。

* 本文系国家自然科学基金面上项目“面向跨语言观点摘要的领域知识表示与融合模型研究”（项目编号：71974202）和国家自然科学基金重大课题“国家安全大数据综合信息集成与分析方法”（项目编号：71790612）研究成果之一。
作者简介：余传明（ORCID：0000－0001－7099－0853），教授，E-mail：yucm@zuel.edu.cn；王曼怡（ORCID：0000－0002－0633－0073），硕士研究生；安璐（ORCID：0000－0002－5408－7135），教授，博士生导师。
收稿日期：2020－02－08 修回日期：2020－04－22 本文起止页码：131－144 本文责任编辑：易飞

2 研究现状

按照研究侧重点的不同，本文从传统的实体关系抽取研究、基于深度学习的实体关系抽取研究和跨语言实体关系抽取研究三方面阐述研究现状。

2.1 传统的实体关系抽取研究

经典的实体关系抽取方法主要分为有监督、半监督、弱监督和无监督 4 类。有监督的实体关系抽取方法将关系抽取任务当作分类问题，根据训练数据设计有效的特征，从而学习各种分类模型，然后使用训练好的分类器预测关系，主要分为基于核函数的方法、基于逻辑回归的方法、基于句法解析增强的方法和基于条件随机场的方法。S. Zhao 等[6]将分词、句法解析和深度依存关系分析三个级别的语法信息用核函数表示，以此克服在某个单一级别上的错误，并使用支持向量机(SVM)在数据集上进行方法的评估。N. Kambhatla[7]使用最大熵模型，将来自文本的各种词汇、句法和语义特征结合在一起，用于语义关系抽取，同时证明了大量信息特征的使用有助于提高模型表现。S. Miller 等[8]运用增强的句法解析合并表示句法和语义信息，进而建立集成模型，解决了线性模型句子处理过程中错误累积传播的问题。A. Culotta 等[9]提出一种集成机器学习模型，该模型能够使用线性链条件随机场学习上下文关系和关联关系模式来抽取实体之间的关系，并将抽取任务(已有的关系模式)和挖掘任务(隐藏的关系模式)集成到一起。上述 4 类有监督方法需要手工标注大量的训练数据，浪费时间精力，因此人们继而提出了基于半监督、弱监督和无监督的关系抽取方法来解决人工标注语料问题。半监督的学习方法主要采用 Bootstrapping 进行关系抽取，该方法首先手工设定若干种子实例，然后迭代地从数据中抽取关系对应的关系模板和更多的实例。S. Brin[10]提出 DIPRE 方案，将互联网作为训练集，计算每个模板的特殊性，筛选比较适合的模板用于下一轮的实体关系抽取。弱监督的学习方法涵盖了试图通过较弱的监督来构建预测模型的各种研究。例如，M. Craven 等[11]在研究从文本中抽取结构化数据、建立生物学知识库的过程中首次提出了弱监督机器学习思想。无监督的学习方法则是利用有相同语义关系的实体对进行关系抽取。例如，T. Hasegawa 等[12]在 ACL 会议上首次提出了一种无监督的命名实体之间的关系抽取方法。

2.2 基于深度学习的实体关系抽取研究

经典方法存在特征提取误差传播问题，极大地影响实体关系抽取的效果。随着近些年深度学习的崛起，学者们逐渐将深度学习应用到实体关系抽取的任务中。根据数据集标注量级的差异，基于深度学习的实体关系抽取任务可分为有监督和远程监督两类，其中有监督的实体关系抽取方法是近年来关系抽取的研究热点。该方法能避免经典方法中人工特征选择等步骤，减少并改善特征抽取过程中的误差积累问题。根据实体识别及关系分类两个子任务完成的先后顺序不同，基于深度学习的有监督实体关系抽取方法可以分为流水线(Pipeline)方法和联合学习(Joint Learning)方法。R. Socher 等[13]提出使用递归神经网络(RNN)来解决实体关系抽取问题。该方法对句子进行了句法解析，能够有效地考虑句子的句法结构信息，但同时该方法无法很好地考虑两个实体在句子中的位置和语义信息。D. J. Zeng 等[14]提出利用词向量和词的位置向量作为卷积神经网络(CNN)的输入，引入了实体和其他词的距离信息，可以很好地把句子中实体的信息考虑到关系抽取中。随后，C. N. D. Santors 等[15]提出了一种新的损失函数的 CNN，采用新的损失函数能够有效提高不同实体关系类型的区分度。A. Katiyar 等[16]首次将注意力机制 Attention 与循环神经网络 Bi-LSTM 一起用于联合提取实体和分类关系，神经网络模型在有监督领域的拓展皆取得不错效果。同时基于深度学习的远程监督实体关系抽取方法因具有缓解远程监督数据集中错误标签和特征抽取误差传播问题的能力而成为研究热点，主要基础方法包括 CNN、RNN、LSTM 等网络结构。近年来，学者们在基础方法之上提出了多种改进，如：D. J. Zeng 等[17]在远程监督上采用分段最大池化的分段卷积神经网络(PCNN)，通过分段最大池化层来自动学习相关特征；Y. K. Lin 等[18]在远程监督上提出将 CNN 和注意力机制结合起来使用，使用 CNN 作为句子编码器，并使用句子级别的注意机制。此外，G. L. Ji 等[19]提出在 PCNN 和 Attention 的基础上添加实体的描述信息来辅助学习实体的表示，X. Ren 等[20]提出的 COTYPE 模型、Y. Y. Huang[21]提出的残差网络皆增强了实体关系抽取的效果。

随着实体关系抽取方法的不断优化，学者们逐渐将关系抽取任务应用到学术、农业、医学等不同领域中。蒋婷等[22]利用学术文献的结构特点，在本体概念抽取的基础上，对文献中概念的类型进行分类；俞琰等[23]提出基于依存句法分析的中文专利术语选取方法，能够有效提高中文专利术语抽取的准确性；吴粤敏等[24]使用农业上市公司年报数据，采用基于双重注意

力机制的门控循环单元算法研究中文文本关系的自动抽取；朱慧等[25]构建了面向汉语领域的术语非分类关系抽取模型，引入共现分析、结构分析、模板构建、逻辑推理等方法，为术语非分类关系抽取提供了新的思路；张琴等[26]选取词嵌入表示级别、词汇级别和语法级别的三种类型特征，主要探讨词嵌入表示特征在关系抽取中的作用；陈果等[27]通过融合领域元知识和词嵌入向量类别，使用少量领域知识对心血管等细分领域进行实体关系抽取。

2.3 跨语言实体关系抽取研究

近年来，针对跨语言关系抽取的方法大多以单语言关系抽取为基础，使用机器翻译工具、转移学习方法以及计算机视觉中的生成对抗网络，将多种语言的信息联系起来。现有的大多数研究都试图利用平行数据或基于知识的系统将有效信息从标注资源丰富的语言转换为标注语料相对稀缺的语言。L. H. Qian 等[28]提出了一种基于伪平行语料库和实体对齐的中英文关系分类的双语主动学习模型，实验结果表明用于关系分类的双语主动学习明显优于单语主动学习。S. Kim 等[29]提出了一种使用平行语料库进行关系检测的跨语言注释投影策略，为缺乏标注语料的低资源语言建立了关系抽取系统。胡亚楠等[30]为了充分利用多种语言之间的互补性，提出一种双语协同训练的关系分类方法，可以同时提高每种语言的关系分类性能。M. Faruqu 等[31]采用基于机器翻译的跨语言投影法进行多语言开放关系抽取，它通过使用机器翻译工具将源语言翻译成英语，再对英语句子进行关系抽取，最后将关系短语投影回源语言。P. Verga 等[32]对通用模式关系抽取的范围和灵活性进一步改进，尝试采用多语言转移学习进行多语言关系抽取，但是这些工作是对已存在知识库的语言建立模型，而没有完全利用文本中包含的语义信息。Y. K. Lin 等[33]建立了多语言的远程监督关系抽取数据集，提出基于跨语言注意力机制的神经关系抽取模型（MNRE），它为不同语言中的每个句子建立句子表示，并利用多语言注意力机制获取多种语言数据间的一致性和互补性。X. Z. Wang 等[34]在过去工作的基础上加入对抗策略，提出了对抗多语言神经关系抽取模型（AMNRE），取得了较好的模型效果。B. W. Zou 等[35]提出一种特征适应的方法用于跨语言关系分类，首先利用机器翻译获得目标语言数据集，再利用生成对抗网络将源语言的特征表示迁移到目标语言。

值得说明的是，目前的研究仍然较多地局限于单语言实体关系抽取，针对跨语言实体关系抽取的研究则大多采用机器翻译获得平行语料，研究者们为了能够在翻译后的实例中找到对应实体的位置，提出基于混合匹配原则的实体对齐、启发式的实体对齐等，但仍然无法避免实体对齐错误的问题，从而影响跨语言实体关系抽取模型的表现。鉴于此，本文尝试使用词典扩展或自学习法得到的共享空间双语词典获取跨语言平行语料，再引入生成对抗网络[36]（GAN），通过对抗特征适应将源语言的特征表示迁移给目标语言进行关系抽取。为了检验模型的有效性，我们将本文方法应用到以 ACE2005 中英双语数据集为基础的英语－中文、中文－英文两种跨语言关系抽取任务上，并进一步探究模型的各个模块变化对关系抽取效果的影响。

3 研究方法

3.1 研究问题

本文旨在探究跨语言情境下的实体关系抽取问题，即在目标语言标注语料缺乏的情况下，通过在源语言和目标语言之间建立桥梁（机器翻译或者双语词典），得到目标语言的训练语料，再运用源语言和目标语言的平行语料信息训练跨语言的实体关系抽取模型，得到目标语言上的关系抽取模型。假定源语言标注训练语料丰富，标注样本集为 S，其中每个样本实例 $s=\{w_1,w_2,\cdots,w_n\}$，实体 $e_1,e_2\in w_i,i=\{1,\cdots,n\}$，两个实体之间的关系为 y；目标语言标注语料缺乏或无已标注样本，输入未标注样本 $t=\{w_1,w_2,\cdots,w_n\}$，模型旨在利用源语言丰富的标注语料预测目标语言句子中实体 e_1 和 e_2 之间的关系。针对跨语言实体关系抽取，本文深入探究以下问题：①在跨语言关系抽取模型中，模型结构对于关系抽取效果是否有显著影响？②在跨语言对抗关系抽取模型训练过程中，如何使用共享空间双语词向量对词嵌入初始化？词嵌入是否微调对模型表现有哪些影响？③在跨语言平行语料获取模块，如何合理地确定源语言和目标语言双语词典的规模？双语词典的规模是否越大越好？④在跨语言实体关系抽取任务中，如何合理地确定源语言和目标语言训练数据的规模？训练数据的规模是否越大越好？⑤与有监督的目标语言实体关系抽取模型相比，无监督模型的表现是否存在显著差异？

3.2 模型描述

针对上述研究问题，本文提出结合跨语言平行语料获取和对抗适应关系抽取的跨语言对抗关系抽取框架（见图 1）。该框架由跨语言平行语料获取（Cross-

Lingual Parallel Corpus Acquisition, CLPCA)模块和对抗适应关系抽取(Adversarial Adaptation Relation Extraction, AARE)模块两部分构成。CLPCA 模块通过词典扩展或自学习法得到共享空间双语词典,在此基础上获取跨语言平行语料。AARE 模块引入生成对抗网络(GAN)。首先,利用源语言句子编码器(Source-Language Sentence Encoder,SSE)和目标语言句子编码器(Target-Language Sentence Encoder,TSE)分别学习源语言实例和上述获得的目标语言实例的潜在特征表示。其次,将它们输入判别器去判别其是来自源语言还是目标语言,通过对抗特征适应将源语言的特征表示迁移给目标语言。最后,利用训练得到的目标语言关系抽取网络对目标语言实例进行关系分类。

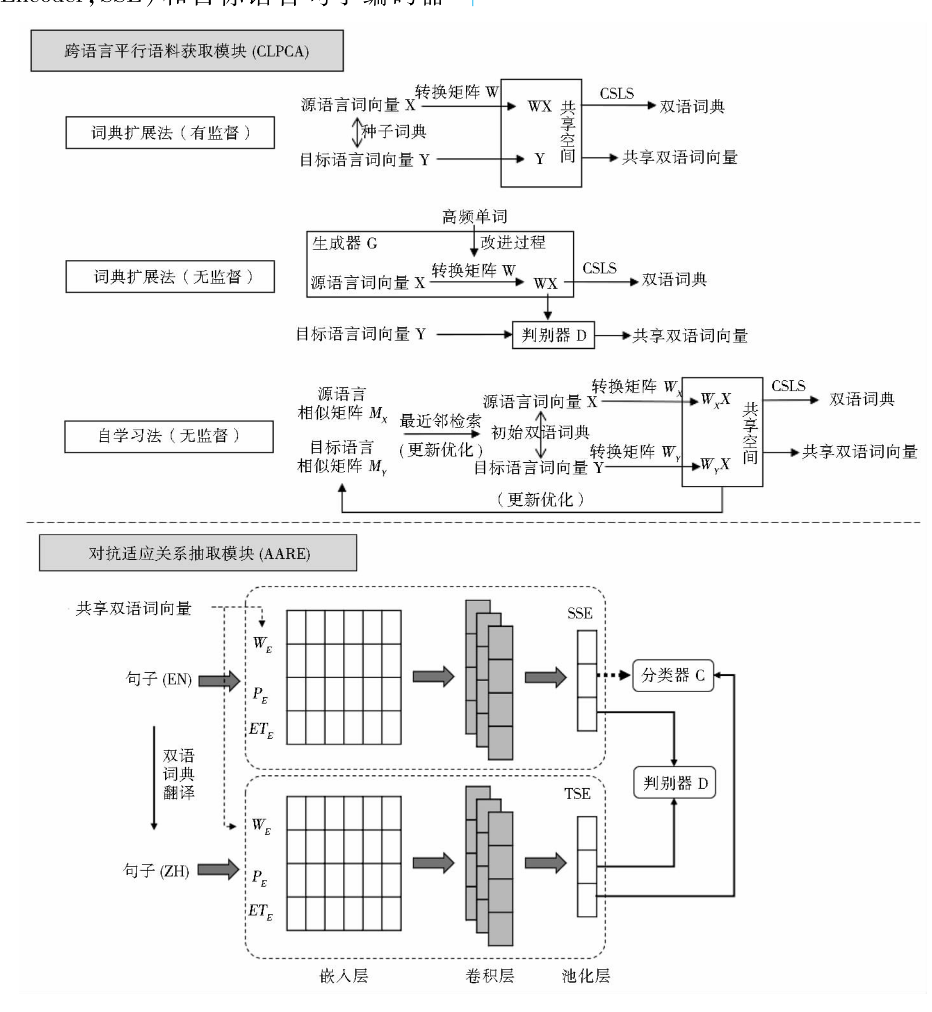


图 1　跨语言对抗关系抽取(CLARE)框架

3.2.1　跨语言平行语料获取模块

最近的研究表明,多数跨语言实体关系抽取研究通过机器翻译的方式,将源语言翻译成目标语言获得平行语料,这种方式获得的平行语料可能由于实体无法对齐而造成实体位置信息错误。考虑到实体位置信息在实体关系抽取任务中的重要性,关系实例中位置信息的错误传播很可能影响关系抽取模型的表现。A. Conneau 等[37]2018 年提出基于双语词典的词典扩展法进行单词翻译,M. Artetxe 等[38]提出无监督的自学习法学习跨语言映射词向量,受其工作的启发,在跨语言情境下的实体关系抽取中,我们将词典扩展法和自学习法应用于双语平行语料的获取。具体来说,本文采用有监督的词典扩展法以及无监督的词典扩展法和自学习法分别获取源语言和目标语言在共享空间中的双语词向量及双语词典,再利用双语词典将源语言的单词翻译成目标语言,并将源语言的关系标签直接映射给目标语言,得到目标语言的训练数据集。此外,由于双语词典的有限性,无法将所有的源语言单词翻译成目标语言,因此获得的目标语言数据集中仍包含少量源语言单词。为了更好地学习源语言和目标语言实例的特征表示,我们使用上述得到的共享空间中的双语词向量初始化词嵌入。

（1）词典扩展法（有监督）。首先通过种子词典学习源语言和目标语言词向量之间的正交性矩阵W，将源语言和目标语言映射到同一向量空间，得到共享空间中的双语词向量；再进行词典规约（Lexicon Induction）[39]，将预训练好的源语言和目标语言词向量通过映射矩阵和跨领域相似度局部缩放（Cross-domain Similarity Local Scaling，CSLS）[37]，得到包含更多单词对的双语词典。有监督的词典扩展法主要分为两个部分：

一是正交性映射。假设有一个种子词典 $D=\{x_i, y_i\}$ $(i=1,2,\cdots,d)$，其中 x_i 为源语言的词向量，y_i 为对应目标语言的词向量，共有d个单词对。通过迭代训练公式（1）得到W为正交矩阵，因此可以保证映射前和映射后的两个词向量之间的夹角不变。

$$\min_W = \| Wx_i - y_i \|^2 \quad s.t.\ WW^T = I \qquad \text{公式(1)}$$

本文使用FastText[40]300维的源语言词向量X和目标语言词向量Y。利用正交性矩阵W，将WX和Y映射到同一个词向量空间。

二是CSLS。当源语言和目标语言的词向量通过正交矩阵W映射到同一空间后，根据最近邻检索找出同一向量空间下，目标语言词向量对应的源语言词向量翻译。计算WX和Y之间的余弦相似度，余弦值越大，说明源语言对应的目标语言翻译越正确。但该方法在高维空间可能存在枢纽点（hubness）问题[41]，即某些点会成为大多数点的最近邻居，因此使用最近邻搜索无法准确找到与每个词语语义最接近的词语，采用CSLS方法惩罚这种枢纽点的相似度分数。对于映射到同一空间中的两个词向量 x 和 y，计算它们之间的CSLS分数作为两个词语之间的最终相似度分数，如公式（2）所示：

$$CSLS(x,y) = 2\cos(x,y) - sim_k(x) - sim_k(y) \qquad \text{公式(2)}$$

在公式（2）中，$\cos(x,y)$ 表示两词语的余弦相似度，$sim_k(x)$ 和 $sim_k(y)$ 分别表示 x 和 y 与其在同一空间中的 k 个最近邻居的余弦相似度均值，作为两个惩罚项以解决高维空间存在的枢纽点问题。

（2）词典扩展法（无监督）。上述有监督的词典扩展法需要学习一个种子词典得到映射矩阵W，但是在大多数情况下双语种子词典难以获得。无监督的词典扩展法能够在没有种子词典的情况下通过对抗的方式学习初始映射矩阵W，并通过普式分析法改进该映射，最后利用CSLS方法得到双语词典。无监督的词典扩展法主要分为3个部分：

一是领域对抗。假设 $X=\{x_1,\cdots,x_n\}$ 和 $Y=\{y_1,\cdots,y_m\}$ 分别为源语言和目标语言的词向量集，从 $WX=\{Wx_1,\cdots,Wx_n\}$ 和Y中通过随机采样训练判别器，希望判别器能够尽可能正确判别样本来源于WX还是Y，同时训练映射矩阵W使得WX和Y尽可能相似，阻碍判别器做出正确判别。

将判别器参数定义为 θ_D，$P_{\theta_D}(source=1 \mid z)$ 表示判别器判别向量z属于源语言词向量映射的概率。判别器的损失函数如公式（3）所示，映射矩阵W的损失函数如公式（4）所示。

$$L_D(\theta_D \mid W) = -\frac{1}{n}\sum_{i=1}^{n} logP_{\theta_D}(source=1 \mid Wx_i) - \frac{1}{m}\sum_{i=1}^{m} logP_{\theta_D}(source=0 \mid y_i) \qquad \text{公式(3)}$$

$$L_W(W \mid \theta_D) = -\frac{1}{n}\sum_{i=1}^{n} logP_{\theta_D}(source=0 \mid Wx_i) - \frac{1}{m}\sum_{i=1}^{m} logP_{\theta_D}(source=1 \mid y_i) \qquad \text{公式(4)}$$

按照I. Goodfellow等提出的生成对抗网络[36]的标准训练过程训练模型，对于每一个输入样本，通过随机梯度更新连续训练判别器和映射矩阵W来最小化损失函数 L_D 和 L_W。

二是改进过程。通过对抗训练获得的矩阵W虽然具有较好的表现，但仍与有监督方法有一定差距。对抗方法试图使所有单词对齐，而不考虑单词的频率。但是，低频词向量更新较少，且更有可能出现在每个语料库的不同上下文中，这使它们更难以对齐。因此在线性映射的假设下，最好仅使用高频单词来推断全局映射。

为了进一步改进学到的映射，我们使用在对抗训练中学到的映射矩阵W来构建合成词典。具体来说，考虑使用频率最高的单词，并仅保留彼此最近的邻居以确保获得高质量的词典。随后，将普式分析法应用于此生成的词典中，考虑到用该算法可以生成更准确的词典，因此迭代地应用此方法。由于使用对抗训练获得的合成词典已经很强大，在进行多次迭代后只会观察到较小的改进。

三是CSLS。在得到改进的映射矩阵W后，将双语词向量映射到同一空间，再使用有监督词汇扩展法中同样的CSLS方法得到更完善的双语词典。

（3）自学习法（无监督）。假设给定词汇表中所有词语间的相似度矩阵，每个词语的相似度值的分布不同，且不同语言中对应的两个词语应具有相同的分布。基于这一假设，能够得到一个初始的匹配词语对词典

D,基于该词典学习不同语言词向量间的映射矩阵,进一步使用一种自学习方法迭代地优化双语词典,最终得到最优的映射矩阵。无监督的自学习法主要分为 3 个部分:

一是构造初始解。由于词向量矩阵 X 和 Z 不对齐,即两种语言词汇表中对应的词语不对齐且词向量在同一维度上也不对齐,因此两种语言之间没有直接的对应关系。首先构造两个替代的词向量矩阵 X' 和 Z',使其在第 j 维上是对齐的,即满足 $X'_{*j} = Z'_{*j}$,基于替换后的部分对齐的两种语言的词向量矩阵构造初始双语词典。为得到替代矩阵 X' 和 Z',使用相似性矩阵来代替词向量矩阵,两种语言的相似性矩阵如公式(5)所示:

$$M_X = XX^T,\ M_Z = ZZ^T \qquad \text{公式(5)}$$

从公式(5)可知,M_X 和 M_Z 分别相当于词向量矩阵 X 和 Z 中行和列的全排列,这种全排列能够用以寻找两种语言之间的词典,即能够通过尝试所有行和列之间的索引排列来寻找 M_X 和 M_Z 之间的最佳匹配。为了避免全排列过程出现组合爆炸问题,首先对相似性矩阵 M_X 和 M_Z 的每一行中的值进行排序,作为替代矩阵。因此,给定一个词语及其在排序后的相似性矩阵中对应的行,能够通过最近邻检索(Nearest Neighbor Retrieval)[42]在另一种语言排序后的相似性矩阵中找到其对应的翻译。最终该过程得到的两个替代矩阵 X' 和 Z' 作为自学习步骤的初始解。

二是自学习。自学习过程主要包括两个步骤,不断重复这两个步骤直至模型收敛,具体步骤如下:

首先通过最大化当前双语词典 D 中不同语言词语之间的相似度 D_{sim} 以得到最优的映射矩阵,相似度如公式(6)所示。在公式(6)中,当目标语言的第 j 个词语是源语言第 i 个词语的翻译,则 $D_{ij} = 1$,否则其值为 0。

$$D_{sim} = \sum_i \sum_j D_{ij}((X_{i*} W_X) \cdot (Z_{j*} W_Z)) \qquad \text{公式(6)}$$

再根据上述得到的映射矩阵计算词向量相似矩阵,进而优化当前双语词典 D。

三是鲁棒性改进。由于最终优化目标与初始的双语词典无关,因此为了避免算法陷入局部最优,通过构造初始解作为自学习过程的初始输入,能够从一定程度上缓解局部最优问题。为了进一步解决该问题从而提高模型的鲁棒性,进一步采用以下改进方式:

随机词典生成:为了促进生成双语词典过程中对搜索空间更多地进行探索,以一定概率 p 随机保留相似矩阵中的部分元素,并将其余元素设置为 0,使得整个的双语词典生成过程是随机的。在训练过程中,根据目标函数的优化情况不断增加概率 p 的值以找到最优的参数设置。

双向词典生成:当基于源语言词语搜索与之对应的目标语言词语从而生成双语词典时,并非所有的目标语言词语最终都能出现在双语词典中,同时有些词语可能会在词典中重复出现。反过来基于目标语言词语寻找对应的源语言词语时同样存在该问题。为充分利用源语言和目标语言中的词语,将分别从上述两个方向生成的双语词典连接起来,消除其中重复的词语对以形成最终的双语词典。

3.2.2 对抗适应关系抽取模块

在对抗适应关系抽取模块,首先运用 CLPCA 模块产生的双语词典对源语言关系抽取数据集进行翻译,得到目标语言数据集。B. W. Zou 等[35] 2018 年运用机器翻译和双语实体对齐获取平行语料,并引入生成对抗网络进行关系抽取,考虑到机器翻译后的双语实体对齐过程准确率较低,因此我们采用词典扩展法和自学习法获取双语平行语料,再通过对抗特征适应进行跨语言的实体关系抽取。具体来说,先利用两个句子编码器分别学习两种语言实例的潜在特征表示,并通过对抗特征适应将源语言的特征表示迁移给目标语言,最终得到具有语言适应性的目标语言关系抽取网络对目标语言进行关系预测。

如图 1 所示,对抗适应关系抽取模块主要由两部分组成。第一部分是句子编码器部分,包括源语言句子编码器(SSE)和目标语言句子编码器(TSE),主要利用 CNN 或 LSTM 对源语言和词典翻译得到的目标语言实例进行特征抽取,将包含实体对的句子实例转换成分布式的潜在特征表示。第二部分是对抗训练部分,将句子编码器 SSE 和 TSE 作为生成器,生成两种语言实例的特征表示作为判别器 D 的输入,然后迭代训练判别器和生成器。判别器尽可能正确地判别输入的特征表示来源于哪种语言,生成器则尽可能生成判别器无法准确分辨的特征表示,同时保证关系分类尽量准确。通过生成器和判别器之间的竞争,最终使得训练得到的目标语言句子编码器具有语言适应性。以下进行详细论述。

(1)句子编码器。我们以 CNN 网络结构的句子编码器为例介绍,使用源语言关系抽取数据集构建 SSE,使用词典翻译得到的目标语言关系抽取数据集构建 TSE,具体包括以下 3 层:

一是嵌入层。借鉴 D. J. Zeng 等[14] 的工作，首先构建嵌入层用实值向量来编码单词、单词位置信息和实体类型。输入句子表示为 $x = \{w_1, w_2, \cdots, w_n\}$，使用词向量矩阵 $W_E \in R^{d_w \times |V|}$ 将每个单词初始化为维度为 d_w 的实值向量，其中 V 表示固定大小的词汇表。由于 SSE 和 TSE 网络结构相同，我们使用 3.2.1 跨语言平行语料获取模块得到的双语词向量进行初始化，使得来自不同语言的单词被映射到同一特征空间中。

在关系抽取任务中，靠近目标实体的单词通常更能够决定实体对之间的关系，因此为了捕捉句子中每个单词与两个实体之间的位置信息，我们将单词分别到头实体和尾实体的相对距离转换为实值向量作为单词的位置嵌入。例如，在句子“Bill_Gates is the founder of Microsoft.”中，单词“founder”到头实体“Bill_Gates”和尾实体“Microsoft”的相对距离分别为 3 和 2。我们使用位置向量矩阵 $P_E \in R^{d_p \times |D|}$ 将相对距离映射为两个维度为 d_p 的实值向量，其中 D 表示相对距离集合。因此，对于每个单词，均能够获得关于两个实体的两个位置向量。

此外，为了能够反映实体类型与实体间关系类型之间的关系，我们对句子中的每个单词加入两个实体的实体类型嵌入，使用向量矩阵 $ET_E \in R^{d_{et} \times |E|}$ 将实体类型映射为两个维度为 d_{et} 的实值向量，其中 E 表示实体类型集合。

最终，我们将一个输入句子表示成一个向量序列 $w = \{w_1, w_2, \cdots, w_n\}$，其中每个单词的嵌入维度为 $d = d_w + 2d_p + 2d_{et}$。

二是卷积层。在编码输入句子后，卷积层使用多个卷积核在句子上滑动来提取局部信息，第 i 个滑动窗口的输出为：

$$p_i = W_c \, w_{i-w+1:i} + b \qquad \text{公式(7)}$$

在公式(7)中，$w_{i-w+1:i}$ 定义为第 i 个窗口内 w 个单词的词向量连接，$W_c \in R^{d_c \times (w \times d)}$ 是卷积矩阵，$b \in R^{d_c}$ 是偏置向量，其中 d_c 表示卷积核的数量，也是卷积层的输出维度。

三是最大池化层。我们利用最大池化层合并卷积层提取的所有局部特征，并应用激活函数 tanh，获得固定长度的最终表示。输出向量 $x \in R^{d_c}$ 的第 j 个元素为：

$$[x]_j = \tan \max_i \mathrm{p}_{ij} \qquad \text{公式(8)}$$

(2)对抗训练。得到句子编码器 SSE 和 TSE 生成的特征表示后，将其输入关系抽取器 C 进行实体关系抽取。关系抽取器 C 由一层全连接和一个 softmax 分类器组成，最后输出每个输入样本在所有关系上的概率分布。我们将句子编码器 SSE 和 TSE 作为生成器，判别器使用一层全连接神经网络和一个 sigmoid 激活函数构建二元分类器，它接收生成器的输出作为输入，判别特征表示来自 SSE 还是 TSE。

具体的对抗训练过程如下：①首先，利用源语言关系抽取数据集预训练 SSE 和关系分类器 C，最小化关系分类损失(公式 9)；②然后训练判别器 D，最小化判别器损失(公式 10)，再在目标语言上训练 TSE 和关系分类器，最小化综合损失函数(公式 13)，同时不断迭代过程②直到模型收敛；③最后，如果判别器无法正确分辨输入特征属于哪种语言，则表明来自 TSE 的输入特征具有了语言适应性。在成功训练后，生成器输出的特征表示既能够进行准确的关系抽取，同时对于不同语言之间，特征表示的差异大大减小。下面具体介绍对抗训练过程中模型的损失函数，主要包括 3 类损失：

一是源语言上的关系分类损失。定义 SSE 和关系分类器 C 的参数分别为 θ_s 和 θ_C，训练目标是最小化交叉熵损失函数，如公式(9)所示：

$$L_{rc\text{-}sse}(\theta_s, \theta_C) = E_{(x_s, y) \sim data}[J(C(H_{sse}(x_s; \theta_s); \theta_C), y)] \qquad \text{公式(9)}$$

在公式(9)中，$L_{rc\text{-}sse}$ 表示源语言上的关系分类损失，$E_{(x_s, y) \sim data})[\cdot]$ 表示对数据分布的期望，$J(p, y)$ 为预测概率分布 p 与真实标签 y 之间的交叉熵损失函数，$C(H_{sse}(x))$ 表示输入 SSE 的特征表示为 $H_{sse}(x)$ 时关系分类器 C 的最终预测，(x_s, y) 为模型的输入和输出，其中 x_s 代表源语言样本实例，y 为关系标签。

二是对抗损失。对抗损失 L_{adv} 用来训练判别器正确判别特征表示来自源语言还是目标语言，定义判别器 D 的参数为 θ_D，判别器的训练目标是尽可能正确地判别特征表示来源，损失函数如公式(10)所示：

$$\min_{\theta_D} L_{adv} = E_{(x_s, x_t, y) \sim data}[\log(1 - D(H_{sse}(x_s; \theta_D))) + \log D(H_{tse}(x_t; \theta_D))] \qquad \text{公式(10)}$$

在公式(10)中，$D(H)$ 为判别器评估特征表示 H 来自 SSE 还是 TSE 的概率输出，H_{sse} 和 H_{tse} 分别表示 SSE 和 TSE 输出的特征表示，且 x_s 代表源语言样本实例，x_t 代表词典翻译得到的目标语言样本实例。

三是目标语言上的关系分类损失。定义 TSE 的参数为 θ_t，TSE 的训练目标是最小化判别器正确判别特征来源的概率，损失函数如公式(11)所示，$L_{tse}(\theta_t)$ 表示目标语言上的判别损失。

$$L_{tse}(\theta_t) = E_{x_t \sim data}[\log D(H_{tse}(x_t; \theta_t))] \qquad \text{公式(11)}$$

定义关系分类器 C 的参数为 θ_C，分类器 C 的训练目标是正确地进行关系分类，并最小化交叉熵损失函数，如公式(12)所示，$L_{rc}(\theta_t,\theta_C)$ 表示目标语言上的分类损失。

$$L_{rc}(\theta_t,\theta_C)=E_{(x_t,y)\sim data}[J(C(H_{tse}(x_t;\theta_t);\theta_C),y)] \quad 公式(12)$$

最后，将公式(11)和公式(12)结合起来，最小化联合损失，如公式(13)所示，其中 β 是用于调整判别损失和分类损失之间权重的平衡参数。

$$\min_{\theta_t,\theta_C} L_{rc\text{-}tse}=\beta L_{tse}(\theta_t)+L_{rc}(\theta_t,\theta_C) \quad 公式(13)$$

4 实验与分析

4.1 数据集

为了更好地探究模型性能，考虑到英文和中文的代表性，本文探究"源语言英语－目标语言中文"和"源语言中文－目标语言英文"两种跨语言关系抽取任务。为使研究更具有代表性，选择在跨语言实体关系抽取任务中最为广泛使用的 ACE 2005 中英文关系抽取数据集[44]。该数据集来源为报纸、广播、新闻专线及博客等，为非平行语料，共定义了六大类关系类型，分别为 PHYS、PART-WHOLE、ART、ORG-AFF、PER-SOC 和 GEN-AFF(不包含 Other)。其中，PHYS 表示地理位置关系，PART-WHOLE 表示部分和整体关系，ART 表示物品所属关系，ORG-AFF 表示组织隶属关系，PER-SOC 表示人际交往关系，GEN-AFF 表示居民的宗教种族等隶属关系。在对语料进行了预处理(包括提取文本、分句、中文分词、特征提取等)后，将中英文语料划分成训练集、验证集和测试集，数据集的具体划分情况见表 1。

表 1 ACE 2005 中英文语料的数据描述

关系类型	英文(EN)			中文(CH)		
	训练集	验证集	测试集	训练集	验证集	测试集
PHYS	1 100	278	278	1 192	205	197
PART-WHOLE	775	162	182	1 649	294	336
ART	491	96	151	476	97	59
ORG-AFF	1 472	365	359	1 611	226	359
PER-SOC	438	106	77	465	83	116
GEN-AFF	512	124	104	1 462	270	199
总计	4 788	1 131	1151	6 855	1 175	1 266

4.2 参数设置

在本实验中，跨语言平行语料获取模块使用 Fast-text[45] 预训练的英文和中文词向量，词向量维度为 300 维，具体的模型参数设置见表 2。对抗适应关系抽取模块使用上一模块生成的 300 维、大小为 10 万的共享空间双语词向量初始化模型，具体的模型参数设置见表 3。

表 2 跨语言平行语料获取模块参数设置

模型名称	参数名称	参数值
词典扩展法(有监督)(MUSE_sup)	n_refinement	5
	max_vocab	100 000
	初始双语词典大小	5 000
	扩展双语词典大小	100 000
词典扩展法(无监督)(MUSE_un)	batch_size	32
	epoch 数	5
	迭代数/epoch	1 000 000
	optimizer	SGD
	max_vocab	100 000
	初始双语词典大小	0
	扩展双语词典大小	100 000
自学习法(无监督)(Vecmap)	batch_size	10 000
	max_vocab	100 000
	初始双语词典大小	0
	扩展双语词典大小	100 000

表 3 对抗适应关系抽取模块的参数

	参数名	参数值
I/O	word_emb_dim	300
	pos_emb_dim	20
	entype_emb_dim	30
Training Setting	iteration	100
	shuffle	True
	emb_update	False
	optimizer	Adadelta
	batch_size	100
Hyperparameters	kernel_size	3
	kernel_num	100
	cnn_dropout	0.5
	lstm_hidden_dim	100
	lstm_layer	1
	lstm_dropout	0
	bilstm = True	True
	ρ	0.9
	β	0.5

4.3 评价指标

由于跨语言实体关系抽取任务的关系标签数为多个，属于多分类任务，因此我们采用 Macro-P(简写为 P)、Macro-R(简写为 R)、Macro-F1(简写为 F1)以及精确率 Accuracy 作为评价指标。其中，Macro-P、Macro-R 和 Macro-F1 表示在多分类任务中的每个类别上分别计算准确率 P、召回率 R 和 F1 值，再将多个类别下计

算得到的值进行平均,得到总体值;精确率 Accuracy 表示预测正确的样本占总样本的比率。

4.4 基础实验结果

在基础实验中,任务 1 以英语为源语言、中文为目标语言(简写为 EN⇒CH);任务 2 以中文为源语言、英语为目标语言(简写为 CH⇒EN)。在跨语言平行语料获取模块(CLPCA),通过训练得到的双语词典将资源丰富的源语言数据集翻译为目标语言,同时使用得到的共享空间中的双语词向量对下一模块关系抽取模型的输入样本单词进行初始化。在对抗适应关系抽取模块(AARE),用源语言训练集和词典翻译得到的目标语言训练集,训练跨语言对抗关系抽取模型,得到测试集上的实验结果。我们将本文模型与 B. W. Zou 等[35]提出的基于对抗学习关系抽取方法及胡亚楠等[30]提出的基于平行语料获取的关系抽取方法进行对比。除此之外,我们对目标语言进行单语言的实体关系抽取,并将其实验结果作为我们的模型上限进行对比。具体实验结果如表 4、表 5 所示。

表 4 跨语言对抗关系抽取实验结果(EN⇒CH)

Model (EN⇒CH)	Accuracy	P	R	F1
CNN-CH	**0.903 6**	0.916 4	0.909 4	**0.912 9**
传统的对抗学习方法(B. W. Zou 等[35])	/	0.687 3	0.723 5	0.705 0
传统的平行语料获取方法(胡亚楠等[30])	/	0.813 0	0.812 0	0.812 0
MUSE_sup + CNN + GAN	**0.863 3**	0.878 1	0.882 0	**0.880 1**
MUSE_un + CNN + GAN	0.797 8	0.855 5	0.810 5	0.832 4
Vecmap + CNN + GAN	0.853 9	0.858 3	0.856 3	0.857 3
MUSE_sup + LSTM + GAN	**0.857 0**	0.870 7	0.862 8	**0.866 7**
MUSE_un + LSTM + GAN	0.812 8	0.845 6	0.834 7	0.840 1
Vecmap + LSTM + GAN	0.843 6	0.839 5	0.853 6	0.846 5

表 5 跨语言对抗关系抽取实验结果(CH⇒EN)

Model (CH⇒EN)	Accuracy	P	R	F1
CNN-EN	**0.911 4**	0.897 1	0.903 7	**0.900 4**
传统的对抗学习方法(B. W. Zou 等[35])	/	0.695 1	0.737 4	0.715 6
传统的平行语料获取方法(胡亚楠等[30])	/	0.801 0	0.798 0	0.799 0
MUSE_sup + CNN + GAN	0.820 2	0.817 5	0.835 4	0.826 3
MUSE_un + CNN + GAN	0.731 5	0.786 6	0.783 5	0.785 0
Vecmap + CNN + GAN	**0.841 0**	0.828 4	0.842 6	**0.835 5**
MUSE_sup + LSTM + GAN	**0.845 4**	0.836 0	0.848 5	**0.842 2**
MUSE_un + LSTM + GAN	0.737 6	0.787 0	0.783 7	0.785 4
Vecmap + LSTM + GAN	0.826 2	0.822 9	0.823 9	0.823 4

从表 4 可以看出,在源语言英语-目标语言中文(EN⇒CH)的实体关系抽取任务上,MUSE_sup + CNN + GAN 的模型表现最好,F1 值为 0.880 1,距离中文单语言实体关系抽取上限仅 3.28%。与传统的基于对抗学习的跨语言关系抽取模型(表 4 第 2 行)和传统的基于平行语料获取的跨语言关系抽取模型(表 4 第 3 行)相比,本文的模型表现有较大提升,表明我们提出的 CLARE 框架能够有效地对资源稀缺的语言进行关系抽取,显著提升跨语言实体关系抽取的效果。针对 CLPCA 模块,在三种词典翻译方法(MUSE_sup、MUSE_un 和 Vecmap)中,MUSE_sup 模型的表现最好,在 CNN + GAN 模型下 F1 值比 MUSE_un 和 Vecmap 模型分别高出 4.77% 和 2.28%。同时,有监督模型表现优于无监督模型,说明在词典翻译过程中,种子词典能够帮助模型学习正确的双语映射,更好地进行词典扩展,因此引入适当的外部知识有助于模型训练。在 CLPCA 模块采用相同方法时,比较 AARE 模块的网络结构可以发现:在多数情况下(MUSE_sup 和 Vecmap),CNN 的表现优于 LSTM,F1 值分别高出 1.34% 和1.08%;在少数情况下(MUSE_un),CNN 表现略低于 LSTM。

从表 5 可以看出,在源语言中文-目标语言英语(CH⇒EN)的实体关系抽取任务上,MUSE_sup + LSTM + GAN 的模型表现最好,F1 值为 0.842 2,与英文单语言实体关系抽取上限相差 5.82%。与传统的基于对抗学习的跨语言关系抽取模型(表 5 第 2 行)和传统的基于平行语料获取的跨语言关系抽取模型(表 5 第 3 行)相比,本文的模型表现同样有较大提升,表明了 CLARE 框架在跨语言实体关系抽取任务上的有效性。同时,CLPCA 模块 MUSE_sup 和 Vecmap 的模型表现比较接近,MUSE_un 模型表现最差,它的 F1 值远远低于 Vecmap 模型,表明无监督模型中 Vecmap 的无监督效果优于 MUSE_un 模型。此外,同样在 CLPCA 模块的方法相同时,比较 AARE 模块的网络结构可以发现:在目标语言为中文时,多数情况下(MUSE_sup 和 MUSE_un),LSTM 的表现优于 CNN,F1 值分别高出 1.59% 和 0.04%,在 Vecmap 模型下,CNN 表现更好。

4.5 扩展实验结果

在扩展实验部分,探究模型结构、共享空间词向量微调、不同大小的双语词典进行词典翻译以及不同大小的训练数据集对跨语言实体关系抽取效果的影响,同时将模型结果与不同大小数据集的有监督实体关系抽取模型结果进行比较。

4.5.1 模型结构对跨语言实体关系抽取效果的影响

为了探究模型对抗部分对跨语言实体关系抽取效果的影响,我们在 EN⇒CH 任务上固定跨语言平行语料获取模块,将对抗适应关系抽取模块的双语对抗部分去除,仅使用 CNN 模型在词典翻译得到的目标语言训练集上进行关系抽取,比较对抗去除前后在 MUSE_sup、MUSE_un 和 Vecmap 3 种模型上评价指标的变化情况。实验结果如表 6 所示:

表 6 模型结构对实验结果的影响

Model (EN⇒CH)	Accuracy	P	R	F1
MUSE_sup + CNN + GAN	0.863 3	0.878 1	0.882 0	0.880 1
MUSE_sup + CNN	0.861 8	0.875 5	0.868 3	0.871 9
MUSE_un + CNN + GAN	0.797 8	0.855 5	0.810 5	0.832 4
MUSE_un + CNN	0.797 0	0.816 6	0.835 2	0.825 8
Vecmap + CNN + GAN	0.853 9	0.858 3	0.856 3	0.857 3
Vecmap + CNN	0.846 0	0.852 6	0.865 8	0.859 2

从表 6 可以看出,在去除对抗部分之后,在 MUSE_sup 和 MUSE_un 模型上的表现均有一定程度的下降,F1 值分别下降了 0.82% 和 0.66%;而对于 Vecmap 模型,去除对抗之后虽然 F1 值提升了0.19%,但 Accuracy 值下降了 0.79%,考虑到数据不平衡对 F1 值的影响,同样认为模型表现有所下降。总的来说,对抗部分能够提高有监督和无监督的词典扩展法以及无监督的自学习法 3 种方法下的模型表现。

4.5.2 共享空间词向量微调对跨语言实体关系抽取效果的影响

为了促进对抗适应关系抽取模块的模型训练,使得目标语言关系抽取模型具有语言适应性,我们的模型使用前一模块得到的双语共享空间词向量对单词向量进行初始化,并在之后的模型训练过程中保持词向量不变。因此,我们在 EN⇒CH 任务上进行了词向量微调情况下的对比实验,探究共享空间词向量微调对跨语言实体关系抽取效果的影响。实验结果如表 7 所示:

表 7 共享空间词向量微调对实验结果的影响

Model (EN⇒CH, CNN + GAN)	Accuracy	P	R	F1
MUSE_sup + 不变	0.863 3	0.878 1	0.882 0	0.880 1
MUSE_sup + 微调	0.852 3	0.877 2	0.871 2	0.874 2
MUSE_un + 不变	0.797 8	0.855 5	0.810 5	0.832 4
MUSE_un + 微调	0.765 4	0.858 9	0.813 8	0.835 8
Vecmap + 不变	0.853 9	0.858 3	0.856 3	0.857 3
Vecmap + 微调	0.838 1	0.849 7	0.859 7	0.854 7

从表 7 可以看出,总体来说,保持共享空间词向量不变能够较为显著地提高模型表现。具体而言,对于 MUSE_sup 和 Vecmap 方法来说,Accuracy 值分别有 1.10% 和 1.58% 的提高;对于 MUSE_un 模型来说,F1 值有略微下降,但 Accuracy 值提高了 3.24%。这说明训练过程中词向量的微调可能会导致共享空间信息的丢失,从而降低模型表现。

4.5.3 双语词典大小对跨语言实体关系抽取效果的影响

模型的跨语言平行语料获取模块通过生成双语词典,对源语言数据集进行词典翻译,得到目标语言的训练数据集。为了探究双语词典大小对跨语言实体关系抽取效果的影响,我们在 EN⇒CH 任务的 MUSE_sup + CNN + GAN 模型上,使用不同大小双语词典翻译的数据集训练模型。具体实验结果如表 8 所示:

表 8 不同大小的双语词典对实验结果的影响

Model (EN⇒CH, MUSE_sup + CNN + GAN)	Accuracy	P	R	F1
100 000	0.863 3	0.878 1	0.882 0	0.880 1
80 000	0.861 0	0.874 8	0.864 8	0.869 8
60 000	0.849 1	0.861 5	0.866 8	0.864 2
40 000	0.855 5	0.870 3	0.865 4	0.867 8
20 000	0.850 7	0.864 1	0.865 3	0.864 7
10 000	0.845 2	0.853 5	0.865 8	0.859 6

由表 8 可以看出,在 EN⇒CH 任务的 MUSE_sup + CNN + GAN 模型上,双语词典大小分别设置为 100 000、80 000、60 000、40 000、20 000 和 10 000,当双语词典大小为 100 000 时,模型结果最优,F1 值为 0.880 1;当双语词典大小为 10 000 时,模型结果最差,F1 值为 0.859 6。总体来看,随着双语词典大小的增加,F1 值呈整体上升趋势。实验结果表明,双语词典大小越大,包含的跨语言知识也越多,从而词典翻译得到的数据集更准确,跨语言实体关系抽取模型的性能也越好。值得说明的是,当双语词典大小为 60 000 时,模型的 F1 值(0.864 2)较词典大小为 40 000 时(0.867 8)略微有所下降。可能的原因在于,一方面,当双语词典从 40 000 增加到 60 000 时,词典未能有效提升双语词汇的覆盖面(即新增的词汇并未反映在测试数据集中),从而导致结果并无提升;另一方面,由于双语词典的扩大,增加了模型训练的复杂度,从而导致模型效果略有下降。这表明,在跨语言研究的实际应用中,仍需综合考虑训练复杂度及双语词汇覆盖面,将

词典规模控制在合理范围[45－46]。

4.5.4 训练集大小对跨语言实体关系抽取效果的影响

为了探究不同训练集的大小对跨语言实体关系抽取效果的影响,我们在 EN⇒CH 任务的 MUSE_sup + CNN + GAN 模型上,使用不同大小的双语平行语料训练模型。对于每种语言来说,训练语料大小设置为 500－4 500,具体实验结果如图 2 所示:

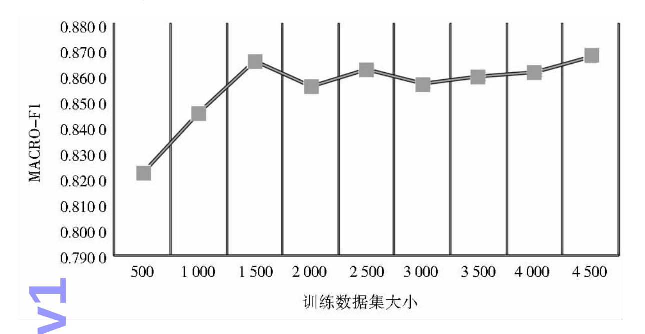


图 2　不同大小的训练数据集对实验结果的影响

由图 2 可知,总体来看,随着训练数据集大小的增加,F1 值呈上升趋势,模型表现在训练数据集大小为 4 500时最好。具体而言,在训练数据集由 500 增加到 1 500的过程中,模型表现有显著提升,但在数据集大小 2 000 时有一定下降,之后由 2 000 增加到 4 500 过程中,F1 值缓慢提升。这表明随着训练数据集的增加,前期模型表现提升较大,后期提升较为缓慢甚至略有下降。

4.5.5 与有监督实体关系抽取模型的结果比较

在本文研究中,我们假设源语言为语料丰富的语言,目标语言语料相对缺乏,在完全无监督情境下,对目标语言进行实体关系抽取。为了与有监督情境下单语言实体关系抽取模型(CNN-CH)进行比较,我们针对 EN⇒CH 任务的 MUSE_sup + CNN + GAN 模型(EN⇒CH, MUSE_sup + CNN + GAN),使用不同大小的中文标注训练集训练。对比实验结果如图 3 所示:

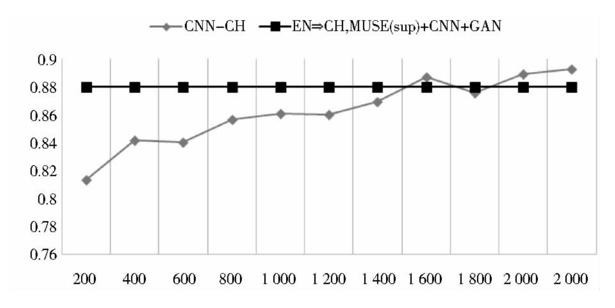


图 3　与有监督实体关系抽取模型的结果比较

从图 3 可以看出,从 200 开始增加中文有监督标注数据,每次增加 200,F1 值逐渐提高且逐渐变缓,直到数据量达到 2 000 时,有监督实体关系抽取的模型表现超过我们的完全无监督模型。这表明我们的完全无监督模型与有监督模型相比,可以取得相对可比的表现,近似相当于 1 800 条标注数据集下有监督模型的性能。

4.6 讨论

根据上述实验结果,我们在 CLARE 框架基础上探究各个子模型的效果。在源语言英语－目标语言中文(EN⇒CH)和源语言中文－目标语言英语(CH⇒EN)的实体关系抽取任务上,总体来说,有监督的词典扩展法的模型表现最好,无监督的词典扩展法的模型表现最差,且无监督模型中自学习法优于词典扩展法。有监督词典扩展法表现优于无监督模型,表明在词典翻译过程中,种子词典能够帮助模型学习正确的双语映射,从而更好地进行词典扩展。

针对问题 1(在跨语言关系抽取模型中,模型结构对于关系抽取效果是否有显著影响?),从模型结构对跨语言实体关系抽取效果的影响可以看出,在模型中加入对抗部分,能够提高有监督和无监督的词典扩展法以及无监督的自学习法 3 种方法下的模型表现。其中,对于自学习法来说,对抗部分带来的提高较小。这表明模型对抗部分对于词典扩展法有更大的帮助,对于已经包含对抗思想的自学习法来说,再次使用对抗并没有取得更好的结果。

针对问题 2(在跨语言对抗关系抽取模型训练过程中,如何使用共享空间双语词向量初始化词嵌入?词嵌入是否微调对模型表现有哪些影响?),从共享空间词向量微调对跨语言实体关系抽取效果的影响可以看出,在训练过程中,保持对抗适应关系抽取模块共享空间词向量不变,能够提高有监督和无监督的词典扩展法以及无监督的自学习法 3 种方法下的模型表现,其中,对自学习法的提高最为明显。这说明训练过程中词向量的微调可能会导致共享空间信息的丢失,使得源语言和目标语言的句子编码更难以映射到同一空间中且位置相近,导致模型性能降低。

针对问题 3(在跨语言平行语料获取模块,如何合理地确定源语言和目标语言双语词典的规模?双语词典的规模是否越大越好?),从不同大小的双语词典对跨语言实体关系抽取效果的影响可以看出,随着双语词典的增大,跨语言对抗关系抽取模型的性能总体上呈现逐渐变好的趋势。由于双语词典越大,包含源语言和目标语言间的信息越多,使用词典翻译将源语言数据集翻译到目标语言也更加准确。在双语词典的规

模达到一定程度后,跨语言实体关系抽取的性能增长缓慢。这表明,在实际应用中,采用规模适度且高度对齐的双语词典即可有效提高跨语言实体关系抽取的效果。

针对问题 4(在跨语言实体关系抽取任务中,如何合理地确定源语言和目标语言训练数据的规模?训练数据的规模是否越大越好?),从不同大小的训练数据集对跨语言实体关系抽取效果的影响可以看出,当双语平行训练语料每种语言大小在 500 – 4 500 之间,随着训练集大小的增加,跨语言实体关系抽取的效果呈现总体上升趋势,且在前期模型表现提高显著,在训练集增加到 1 500 之后,模型表现的提高逐渐趋于平缓甚至有所下降。实验结果表明在特定的阈值下(本文为 1 500 左右),增加训练数据能够更有效地提高跨语言实体关系抽取模型的表现。

针对问题 5(与有监督的目标语言实体关系抽取模型相比,无监督模型的表现是否存在显著差异?),将我们完全无监督的跨语言对抗关系抽取模型与有监督关系抽取模型进行对比,使用不同大小的中文标注语料训练中文实体关系抽取模型。从实验结果可以看出,随着中文标注数据量的增大,中文单语言实体关系抽取效果不断提高且逐渐变缓,直到数据量达到 2 000 时,有监督实体关系抽取的模型表现超过无监督模型。表明无监督模型与有监督模型相比,可以取得相对可比的表现,在本文实验配置下无监督模型与 1 800 条标注数据集下有监督模型的性能接近。

5 结语

为了提升跨语言情境下低资源语言实体关系抽取模型的性能,本文提出了跨语言对抗关系抽取框架,从跨语言平行语料获取和对抗适应关系抽取两个方面进行句子级别的跨语言实体关系抽取。跨语言平行语料获取模块是通过词典扩展或自学习的方法将源语言关系抽取数据集转换为目标语言数据集,解决目标语言数据集缺乏的问题;对抗适应关系抽取模块则是利用对抗特征适应将源语言的特征表示迁移给目标语言,再利用训练得到的目标语言关系抽取网络对目标语言进行关系分类。在"源语言英语 – 目标语言中文"和"源语言中文 – 目标语言英语"两种跨语言关系抽取任务上的实验结果表明,该模型在两种跨语言关系抽取任务上的表现较好,两个任务上最优模型的 F1 值分别为 0.880 1 和 0.842 2,这表明本文提出的跨语言实体关系抽取框架能显著提升低资源语言实体关系抽取的效果。研究结果对于改进跨语言情境下的关系抽取模型,促进实体关系抽取研究在情报学领域的应用具有重要意义。

受制于实验条件,本文的工作还存在一些不足,在后续研究中,我们将开展以下研究:①进一步探究在半监督的情境下,仅以加入部分目标语言标注数据,通过多任务学习或模型融合的方式将目标语言知识与跨语言知识相结合,提高跨语言实体关系抽取的模型表现;②将本文的跨语言系统应用到更多语言情境下的实体关系抽取任务中,解决更多低资源语言的实体关系抽取问题。

参考文献:

[1] 鄂海红,张文静,肖思琪,等. 深度学习实体关系抽取研究综述[J]. 软件学报,2019,30(6):1793 – 1818.

[2] 胡莺夕. 基于深度学习的多实体关系识别及自动文本摘要方法研究与实现[D]. 北京:北京邮电大学,2019.

[3] 郑实福,刘挺,秦兵,等. 自动问答综述[J]. 中文信息学报,2002(6):46 – 52.

[4] 刘峤,李杨,段宏,等. 知识图谱构建技术综述[J]. 计算机研究与发展,2016,53(3):582 – 600.

[5] KOEHN P. A parallel corpus for statistical machine translation[C]//Proceedings of the third workshop on statistical machine translation. Stroudsburg: ACL Press, 2005:3 – 4.

[6] ZHAO S, GRISHMAN R. Extracting relations with integrated information using kernel methods[C]//Proceedings of the annual meeting of the Association for Computational Linguistics. Stroudsburg: ACL Press, 2005: 419 – 426.

[7] KAMBHATLA N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations[C]//Proceedings of the annual meeting of the Association for Computational Linguistics. Stroudsburg: ACL Press, 2004:178 – 181.

[8] MILLER S, FOX H, RAMSHAW L, et al. A novel use of statistical parsing to extract information from text[C]//Proceedings of the 2000 conference of the North American chapter of the Association for Computational Linguistics. Stroudsburg: ACL Press, 2000: 226 – 233.

[9] CULOTTA A, MCCALLUM A, BETZ J T, et al. Integrating probabilistic extraction models and data mining to discover relations and patterns in text[C]//Proceedings of language and technology conference. New York: ITC Press, 2006: 296 – 303.

[10] BRIN S. Extracting patterns and relations from the World Wide Web[C]//Proceedings of international workshop on the Web and databases. Berlin: Springer, 1998: 172 – 183.

[11] CRAVEN M, KUMLIEN J. Constructing biological knowledge bases by extraction information from text sources[C]//Proceedings of the seventh international conference on intelligent systems for molecular biology. Menlo Park: AAAI Press,1999:77 – 86.

［12］HASEGAWA T, SEKINE S, GRISHMAN R. Discovering relations among named entities from large corpora［C］//Proceedings of the annual meeting on Association for Computational Linguistics. Stroudsburg: ACL Press, 2004:415.

［13］SOCHER R, HUVAL B, MANNING C D, et al. Semantic compositionality through recursive matrix-vector spaces［C］//Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. Stroudsburg: ACL Press, 2012: 1201－1211.

［14］ZENG D J, LIU K, LAI S W, et al. Relation classification via convolutional deep neural network［C］//Proceedings of the annual meeting of the Association for Computational Linguistics. Stroudsburg: ACL Press, 2014: 2335－2344.

［15］SANTOS C N D, XIANG B, ZHOU B. Classifying relations by ranking with convolutional neural networks［EB/OL］.［2020－01－01］. https://arxiv.org/pdf/1504.06580.pdf.

［16］KATIYAR A, CARDIE C. Going out on a limb: joint extraction of entity mentions and relations without dependency trees［C］//Proceedings of the annual meeting of the Association for Computational Linguistics. Stroudsburg: ACL Press, 2017:917－928.

［17］ZENG D J, LIU K, CHEN Y B, et al. Distant supervision for relation extraction via piecewise convolutional neural networks［C］//Conference on empirical methods in natural language processing. Stroudsburg: ACL Press,2015:1753－1762.

［18］LIN Y K, SHEN S Q, LIU Z Y, et al. Neural relation extraction with selective attention over instances［C］//Proceedings of the annual meeting of the Association for Computational Linguistics. Stroudsburg: ACL Press, 2016:2124－2133.

［19］JI G L, LIU K, HE S Z. Distant supervision for relation extraction with sentence-level attention and entity descriptions［C］// Proceedings of national conference on artificial intelligence. Menlo Park: AAAI Press, 2017: 3060－3066.

［20］REN X, WU Z Q, HE W Q, et al. CoType: joint extraction of typed entities and relations with knowledge bases［C］//Proceedings of the 26th international conference on World Wide Web. Stroudsburg: ACL Press, 2017:1015－1024.

［21］HUANG Y Y, WANG W Y. Deep residual learning for weakly-supervised relation extraction［C］// Proceedings of the 2017 conference on empirical methods in natural language processing. Stroudsburg: ACL Press, 2017: 1803－1807.

［22］蒋婷,孙建军.学术资源本体非等级关系抽取研究［J］.图书情报工作,2016,60(20):112－122.

［23］俞琰,陈磊,姜金德,等.基于依存句法分析的中文专利候选术语选取研究［J］.图书情报工作,2019,63(18):109－118.

［24］吴粤敏,丁港归,胡滨.基于注意力机制的农业金融文本关系抽取研究［J］.数据分析与知识发现,2019,3(5):86－92.

［25］朱惠,王昊,苏新宁,等.汉语领域术语非分类关系抽取方法研究［J］.情报学报,2018,37(12):1193－1203.

［26］张琴,郭红梅,张智雄.融合词嵌入表示特征的实体关系抽取方法研究［J］.数据分析与知识发现,2017,1(9):8－15.

［27］陈果,许天祥.小规模知识库指导下的细分领域实体关系发现研究［J］.情报学报,2019,38(11):1200－1211.

［28］QIAN L H, HUI H T, HU Y N, et al. Bilingual active learning for relation classification via pseudo parallel corpora［C］//Proceedings of the 52nd annual meeting of the Association for Computational Linguistics. Stroudsburg: ACL Press, 2014:582－592.

［29］KIM S, JEONG M, LEE J, et al. Cross-lingual annotation projection for weakly-supervised relation extraction［J］. Transactions on Asian language information processing, 2014, 13(1):1－26.

［30］胡亚楠,惠浩添,钱龙华,等.基于机器翻译的双语协同关系抽取［J］.计算机应用研究,2015,32(3):662－665.

［31］FARUQUI M, KUMAR S. Multilingual open relation extraction using cross-lingual projection［C］// Poceedings of the 2015 conference of the North American chapter of the Association for Computational Linguistics. Stroudsburg: ACL Press, 2015: 1351－1356.

［32］VERGA P, BELANGER D, STRUBELL E, et al. Multilingual relation extraction using compositional universal schema［C］//Proceedings of the 2016 conference of the North American chapter of the Association for Computational Linguistics. Stroudsburg: ACL Press, 2016: 886－896.

［33］LIN Y K, LIU Z Y, SUN M S. Neural relation extraction with multi-lingual attention［C］// Proceedings of the annual meeting of the Association for Computational Linguistics. Stroudsburg: ACL Press,2017:34－43.

［34］WANG X Z, HAN X, LIN Y K, et al. Adversarial multi-lingual neural relation extraction［C］// Proceedings of the 27th international conference on computational linguistics. Stroudsburg: ACL Press, 2018: 1156－1166.

［35］ZOU B W, XU Z Z, HONG Y, et al. Adversarial feature adaptation for cross-lingual relation classification［C］//Proceedings of the 27th international conference on computational linguistics. Stroudsburg: ACL Press ,2018: 437－448.

［36］GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets［C］// Proceedings of the 27th international conference on neural information processing systems. Montreal: ICONIP Press, 2014: 2672－2680.

［37］CONNEAU A, LAMPLE G, RANZATO M A, et al. Word translation without parallel data［C］// Proceedings of the international conference on learning representations. Vancouver: ICLR Press, 2018.

［38］ARTETXE M, LABAKA G, and AGIRRE E. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings［C］//Proceedings of the 56th annual meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2018: 789－798.

［39］IRVINE A, CALLISONBURCH C. A comprehensive analysis of bilingual lexicon induction［J］. Computational linguistics, 2017, 43(2): 273－310.

[40] JOULIN A, GRAVE E, BOJANOWSKI P, et al. Bag of tricks for efficient text classification[C] // Proceedings of the 15th conference of the European chapter of the Association for Computational Linguistics. Stroudsburg: ACL Press, 2017: 427 – 431.
[41] SHIGETO Y, SUZUKI I, HARA K, et al. Ridge regression, hubness, and zero-shot learning[C] // European conference on machine learning. Switzerland: Springer, 2015: 135 – 151.
[42] PAPADOPOULOS S, BAKIRAS S, PAPADIAS D. Nearest neighbor search with strong location privacy[J]. Proceedings of the VLDB endowment, 2010, 3(1/2):619 – 629.
[43] WALKER C, STRASSEL S, MEDERO J, et al. ACE 2005 multilingual training corpus[EB/OL]. [2020 – 02 – 20]. https://catalog.ldc.upenn.edu/LDC2006T06.
[44] Facebook. Word vectors for 157 languages [EB/OL] [2020 – 03 – 01]. https://fasttext.cc/docs/en/crawl-vectors.html.
[45] 余圆圆,巢文涵,何跃鹰,等. 基于双语主题模型和双语词向量的跨语言知识链接[J]. 计算机科学,2019,46(1):238 – 244.
[46] 李亚超,熊德意,张民. 神经机器翻译综述[J]. 计算机学报,2018,41(12):2734 – 2755.

作者贡献说明:
余传明:论文构思、实验数据获取、论文初稿撰写与修改;
王曼怡:完成基础实验和扩展实验、论文初稿撰写与修改;
安璐:论文构思与修改。

Research on the Model of Adversarial Entity Relation Extraction in Cross-Lingual Context

Yu Chuanming[1] Wang Manyi[2] An Lu[3]
[1] School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan 430073
[2] School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan 430073
[3] School of Information Management, Wuhan University, Wuhan 430072

Abstract: [**Purpose/significance**] From the perspective of entity relation extraction, the knowledge acquisition task in a single language context is extended to a cross-language context, and the relation extraction effect of low-resource languages is improved. [**Method/process**] This paper proposed a Cross-Lingual Adversarial Relation Extraction (CLARE) framework, which decomposed cross-lingual relation extraction into parallel corpus acquisition and adversarial adaptation relation extraction. Through dictionary expansion or self-learning methods, the source language relation extraction data set was converted into the target language data set. On this basis, the feature representation of the source language was transferred to the target language using adversarial feature adaptation, and then the target language relation extraction network obtained by training was used to classify the target language. [**Result/conclusion**] The method in this paper is applied to the English-Chinese and Chinese-English cross-lingual relation extraction task based on the ACE2005 multilingual dataset. The Macro-F1 values of the optimal models on the two tasks are 0.880 1 and 0.842 2 respectively, indicating that the proposed CLARE framework for cross-language adversarial relation extraction can significantly improve the effect of low-resource language entity relation extraction. The research results are of great significance for improving the relation extraction model in the cross-lingual context and promoting the application of entity relation extraction research in the field of information science.

Keywords: cross-lingual information extraction entity relation extraction deep learning generative adversarial network